Statistical mechanics of RNA folding: importance of alphabet size

Ranjan Mukhopadhyay, Eldon Emberly*, Chao Tang, and Ned S. Wingreen
NEC Laboratories America, Inc., 4 Independence Way, Princeton, NJ 08540
(Dated: February 1, 2008)

We construct a base-stacking model of RNA secondary-structure formation and use it to study the mapping from sequence to structure. There are strong, qualitative differences between two-letter and four or six-letter alphabets. With only two kinds of bases, most sequences have many alternative folding configurations and are consequently thermally unstable. Stable ground states are found only for a small set of structures of high designability, *i.e.* total number of associated sequences. In contrast, sequences made from four bases, as found in nature, or six bases have far fewer competing folding configurations, resulting in a much greater average stability of the ground state.

I. INTRODUCTION

RNA plays a central role in molecular biology. In addition to transmitting genetic information from DNA to proteins, RNA molecules participate actively in a variety of cellular processes [1]. Examples are found in translation (rRNA, tRNA, and tmRNA), editing of mRNA, intracellular protein targeting, nuclear splicing of pre-mRNA, and X-chromosome inactivation. The RNA molecules involved in these processes do not code for proteins but act as functional products in their own right. In addition, RNA molecules prepared *in vitro* can be selected to bind to specific molecules such as ATP [2]. In all these cases, the information encoded in the sequence of nucleotide bases of each RNA molecule determines its functional three-dimensional structure. The nucleotide sequence is a kind of genotype, *i.e.*, hereditary information, while the folded three-dimensional structure represents phenotype, the physical characteristics on which natural selection operates. The mapping from genotype to phenotype bears on how biological systems evolve, and RNA folding probably constitutes the simplest example of this mapping [3]. Since early life is believed to have been RNA based [1], RNA folding can provide us with important clues about early life and evolution.

RNA is a polynucleotide chain consisting of the four bases: A, U, G, and C. Complementary base pairs (A-U and G-C) can stack to form "stems" which are helical segments similar to the double helix of DNA. These helices, called secondary structures, are generally arranged in a three-dimensional tertiary structure, stabilized by the much weaker interactions between the helices. Representations of secondary structures are shown in Fig. 1. The energy contributions of secondary and tertiary structures are hierarchical[4], with secondary structures largely determining tertiary folding. Secondary structure is frequently conserved in evolution, and structural homology has been used successfully to predict function [5].

In this paper, we investigate the role of alphabet size in the statistical mechanics and selection of RNA secondary structures. We find pronounced differences between two-letter and four or six-letter alphabets. For sequences constructed with two types of bases, only a small fraction of sequences have thermodynamically stable ground-state structures; these structures are also highly designable, *i.e.*, have a large number of associated sequences. Four and six-letter sequences are much more stable on average, but exhibit no strong correlation between designability and thermodynamic stability. We trace this difference to the greater likelihood of competing, alternatively paired configurations when a two-letter alphabet is used.

For RNA, there already exist algorithms that predict secondary structures [6, 7]. These algorithms are intended to apply to real RNA and, consequently, involve a large number of parameters for the different pairing and stacking combinations. Using one of these algorithms, Fontana *et al.*[8] found a broad distribution of designabilities, *i.e.* number of sequences per structure, after structures were grouped by topology. In this paper, we present, instead, a much simpler model for RNA secondary structure designed to elucidate the role of alphabet size.

The organization of this paper is as follows. In section II, we present a base-stacking model for RNA secondary structure and outline the recursive algorithm used to compute the partition function and ground-state structure. In section III, we employ our model to analyze the stability of folded structures. We find a significant difference in stability between two-letter and four or six-letter sequences due to the greater likelihood of alternative folds in the two-letter case. As a consequence of these alternative folds, in the two-letter case, stability correlates with designability, *i.e.*, total number of sequences associated with a structure. In addition, we find that RNA sequences folding to a given structure form a percolating neutral network. Finally, in section IV, we summarize our main conclusions.

II. BASE-STACKING MODEL

We introduce a base-stacking model for RNA secondary-structure formation. It is known that, within

*Current Address: Center for Physics and Biology, Rockefeller University, New York, NY 10028

a stem of base pairs, the largest energy contribution is the *stacking energy* between two adjacent base pairs (rather than the base-pairing energy itself) and the total energy of the stem is the sum of stacking energies over all adjacent base pairs [4]. A single stack $(i, i+1; j-1, j)$ is defined as two adjacent non-overlapping base pairs (i, j) and $(i+1, j-1)$ where $i+1 < j-1$. For this stack $(i, i+1; j-1, j)$ we assign an energy $-E_s$ if (i, j) and $(i+1, j-1)$ are both complementary Watson-Crick base pairs and zero otherwise. We thus neglect differences in energy between, for example, (A,A;U,U), (A,G;C,U) and (G,G;C,C) stacks. We also neglect energy contributions from isolated base pairs that are not part of a stack, and, consequently, do not include isolated base pairs in the secondary structure.

The largest entropic contribution to an RNA structure comes from stretches of unpaired bases. We incorporate a simplified version of this polymer configurational entropy in our model by associating α degrees of freedom with every unpaired base. Thus, the restricted partition function, corresponding to all micro-states compatible with a given secondary structure is

$$Z_{\rm micro} = \alpha^{n_u} \exp\left[\frac{n_s E_s}{k_B T}\right] \tag{2.1}$$

where n_u is the number of unpaired bases, n_s is the number of stacks, and T is the temperature. The restricted free energy is $F_{\rm micro} = -k_B T \ln Z_{\rm micro} = -E_s n_s - k_B T n_u \ln \alpha$.

In this model, since only complementary base pairs can participate in a stack, only a fraction of possible structures are compatible with any given sequence. However, provided the structure is compatible with the sequence, its restricted free energy is independent of the sequence.

The change in free energy due to the formation of an isolated stack is $-E_s + 4k_B T \ln \alpha$; the first term corresponds to the stacking energy and the second to the loss in configurational entropy (since four bases participate in the stack). For every additional adjacent stack the change in free energy is $-E_s + 2k_B T \ln \alpha$, since only two bases are added to the stack. If, for example, $E_s < 4k_B T \ln \alpha$ but $2E_s > 6k_B T \ln \alpha$ (*i.e.*, $3k_B T \ln \alpha < E_s < 4k_B T \ln \alpha$), then formation of an isolated stack would be unfavorable but formation of a segment consisting of two or more adjacent stacks would be favored by a net decrease in free energy. Thus, for an appropriate choice of parameters, the model correctly provides a nucleation cost to the formation of stems. For this paper we choose $\ln \alpha = 1.5$ and $E_s = 5.5 k_B T$, which are physically motivated and correspond to a nucleation cost for the formation of an isolated stack, with a minimum of two adjacent stacks required to form a stable stem. Our results, however, do not depend sensitively on the choice of these parameters.

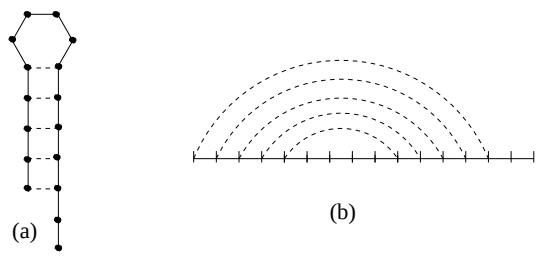

FIG. 1: Representations of RNA secondary structures: (a) flattened-helix diagram of a 16-base structure, and (b) rainbow diagram of the same structure. The restriction that arches do not cross in the rainbow diagram implies the absence of pseudoknots.

In the secondary structure, any two base pairs (i_1, j_1) and (i_2, j_2), with $i_1 < i_2$, are either nested ($i_1 < i_2 < j_2 < j_1$) or independent ($i_1 < j_1 < i_2 < j_2$). Other possibilities correspond to "pseudoknots", which are energetically and kinetically suppressed. It is customary to regard pseudoknots as part of the tertiary structure, and we do not include them here.

In order to compute the ground-state structure and partition function for a given sequence, we make use of the hierarchical nature of secondary structures (due to the absence of pseudoknots). We use a recursive algorithm that is a generalization of the techniques described in Refs. [9] and [10]. Consider the partition function $Z_{i,j}$ for a segment of bases from the position i to $j \geq i$. The base j is either unpaired or can be part of a stack $(k, k+1; j-1, j)$ with $k \in \{i, \ldots, j-3\}$. Thus $Z_{i,j}$ obeys:

$$\begin{aligned} Z_{i,j} &= \alpha Z_{i,j-1} + \sum_{k=i}^{j-3} [Z_{i,k-1} \cdot e^{E_s/k_B T} \\ &\quad \cdot \mathcal{P}_s(k, k+1; j-1, j) \cdot \hat{Z}_{k+2;j-2}], \end{aligned} \tag{2.2}$$

where $\mathcal{P}_s(k, k+1; j-1, j)$ equals 1 if both (k, j) and $(k+1, j-1)$ are complementary base pairs, and equals 0 otherwise; $Z_{i,i-1}$ is defined to equal 1. We have introduced $\hat{Z}_{i,j}$ which is the partition function for the segment with the boundary condition that sites $i-1$ and $j+1$ are paired, implying an energy $-E_s$ for the formation of a bond between the bases at sites i and j. We thus require a second recursion relation for $\hat{Z}$:

$$\begin{aligned} \hat{Z}_{i,j} &= \alpha Z_{i,j-1} + e^{E_s/k_B T} \cdot \mathcal{P}_b(i, j) \cdot \hat{Z}_{i+1,j-1} \\ &\quad + \sum_{k=i+1}^{j-3} [Z_{i,k-1} e^{E_s/k_B T} \mathcal{P}_s(k, k+1; j-1, j) \cdot \hat{Z}_{k+2;j-2}], \end{aligned} \tag{2.3}$$

where $\mathcal{P}_b(i, j)$ equals 1 if (i, j) are complementary base pairs and 0 otherwise. The partition function $Z_{1,N}$ can be computed recursively using (2) and (3) in $O(N^3)$ steps. We use a similar recursive algorithm to compute the ground-state structure $\mathcal{S}_{1,N}$.

III. RESULTS

A. Dependence on Alphabet Size

We have employed our model to analyze the stability of folded structures corresponding to two, four, and six-letter sequences. The thermodynamic stability is defined as the probability $P_{\rm GS}$ that the sequence will be found in the ground state, $P_{\rm GS} = e^{-F_{\rm GS}/k_BT}/Z$ where $F_{\rm GS}$ is the free energy associated with the ground state. Fig. 2 shows a histogram of stability for 40-nucleotide long sequences with ground states containing 12 to 15 stacks [11]. We find four-letter sequences considerably more stable on average than two-letter sequences.

What is the origin of the difference in stabilities between two-letter and four-letter RNA sequences? In order to address this question, we classify the excited-state structures as (i) those formed by breaking existing pairs, and (ii) those formed by re-pairing, *i.e.*, by forming new pairs in addition to breaking existing pairs. Independent of alphabet size, all sequences folding into a given secondary structure $\mathcal{S}$ have the same set of "pair-breaking" excited states. The *sequence* dependence of stability for a given ground-state structure results entirely from re-pairings. *The crucial difference between two-letter and four sequences lies in the substantially greater likelihood of "re-paired" excited states for two-letter sequences.* This follows because the number of pairs one can form in a random sequence of two letters is typically much larger than for a four or six-letter sequence of the same length. For example, for a random four-letter sequence of length N, the probability of forming a stem involving sites i to $i+l$ and $j-l$ to j is lower by a factor of 2^l as compared to a random two-letter sequence of the same length. For the same reason, the fraction of sequences that have highly stacked ground states is much greater for two-letter sequences than for four-letter sequences, and much greater for four than for six.

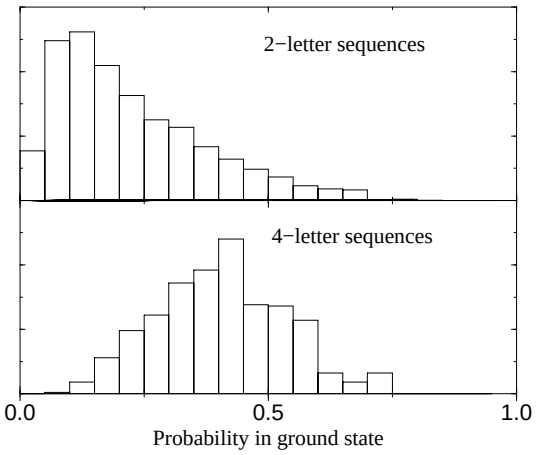

FIG. 2: Histogram of stabilities for: (a) 40-nucleotide long two-letter sequences, and (b) 40-nucleotide long four-letter sequences.

To demonstrate the importance of "re-paired' excited states, we first calculate a "pair-breaking" stability $P_{\rm max} = e^{-F_{\rm GS}/k_BT}/\mathcal{Z}$ where $\mathcal{Z}$ is a pair-breaking partition function calculated by considering only pair-broken excited states. $P_{\rm max}$ gives us an upper bound to the true stability, *i.e.*, probability in ground state $P_{\rm GS}$, that includes competition from re-paired states. In Fig. 3, we plot the true average stability $\langle P_{\rm GS}\rangle$ against the pair-breaking stability $P_{\rm max}$ for two, four, and six-letter sequences. As expected, the average stability is much closer to the maximum set by pair breaking in the case of four-letter sequences than in the case of two-letter sequences. Thus, structures constructed with four-letter sequences are typically much more than stable than those constructed with two letters, and six-letter sequences are typically more stable than four-letter ones. For folding *kinetics*, it is these same "re-paired" states that act as kinetic traps. Due to the lower likelihood of such states, we expect four and six-letter sequences to typically fold faster than two-letter sequences.

B. Stability and Designability

What determines the average stability, $\langle P_{GS}\rangle$, of two-letter sequences? We have seen that for four and six-letter sequences the average stability is close to the "pair-breaking" stability which is determined largely by the

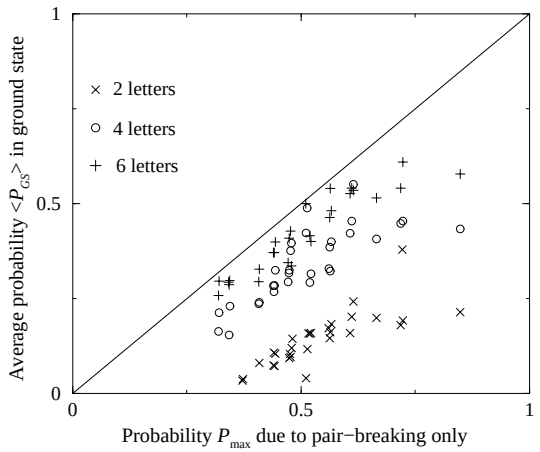

FIG. 3: Average of actual probability in ground state vs. upper bound $P_{\rm max}$ allowing only for breaking of base pairs, plotted for 40-nucleotide sequences. +'s denote RNA sequences constructed from two types of bases, ∘'s denote those constructed from four, and ×'s denote sequences constructed from six types of bases. The actual probability is averaged over sequences with the same pair-breaking stability $P_{\rm max}$ (which is sequence independent).

number of stems and loops. Insight into the stability of two-letter RNA sequences comes from results in protein folding [12, 13, 14, 15, 16, 17]. Based on solvation models with differing hydrophobicities of amino acids, a principle of designability has emerged for protein folding. The designability of a structure is measured by the number of sequences folding uniquely into that structure. A small class of protein stuctures emerge as being highly designable; remarkably, the same class of structures are highly designable whether two or all 20 amino-acid types are used[17]. In a wide range of protein models, sequences associated with highly designable structures are thermodynamically more stable[13, 18] and fold faster than typical sequences[15]. This connection between the designability of a structure and the stability of its associated sequences is referred to as the designability principle. The designability principle reflects a competition among structures. In solvation models, sequences will fold to structures which best match their hydrophobic amino acids to buried sites in the structure (shielded from water). Highly designable structures are those with unusual patterns of surface exposure, and therefore few competitors. This lack of competitors also implies that the sequences folding to such structures are thermally stable. We will now show that the designability principle also holds for two-letter RNA.

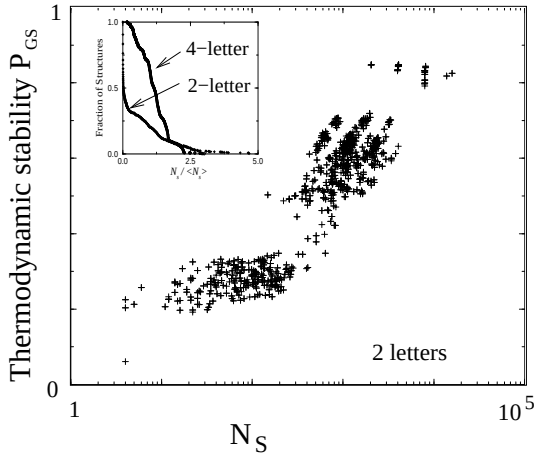

FIG. 4: Stability [20] versus designability N_s (in logarithmic scale) for 24-base RNA sequences constructed with two types of bases. In the inset we plot fraction of compact structures [19] with designability above N_s versus N_s for two and four-letter RNA sequences.

For two base-types (say A and U), we enumerate all sequences and structures of length 24. We find that secondary structures differ considerably in their designability; there are highly designable structures which are ground states of a large number of sequences, and there are poorly designable structures which are ground states of only a few sequences (cf. Fig. 4 inset). In this respect, the results for two-letter sequences are similar to those for protein models[13, 14]. However, the histogram is more noisy for RNA than it is for proteins; so we plot the integrated distribution of designabilities. The most designable structure consists of a stem with a hairpin loop, and a dangling end. We have also studied longer sequences, of lengths 40 and 50, for which we sample sequence space. For 40-nucleotide sequences, the most designable structures consist of a single hairpin loop and dangling ends; a number of double hairpin structures are also highly designable (Fig. 5). For sequences of length 50, double hairpin structures emerge as the most designable. Finally, we find a pronounced correlation between designability and stability of RNA structures. This is shown in Fig. 4 [21] for 24-nucleotide sequences. Thus, two-lette RNA sequences which fold into highly designable secondary structures are unusually thermally stable, verifying the designability principle.

In contrast, for four-letter sequences the range of designabilities is narrower and there is only a weak correlation between designability and stability, with highly stable sequences existing for structures of both high and low designability (Fig. 6). The results for six letters are

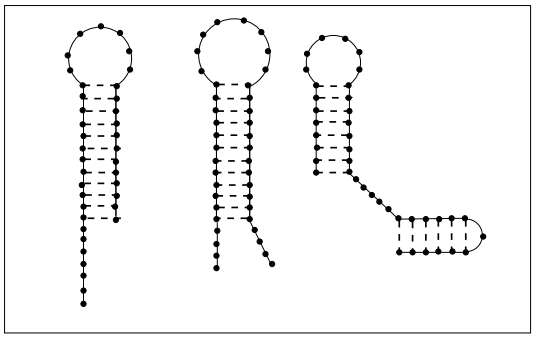

FIG. 5: A few highly designable structures for 40-nucleotide long two-letter RNA sequences.

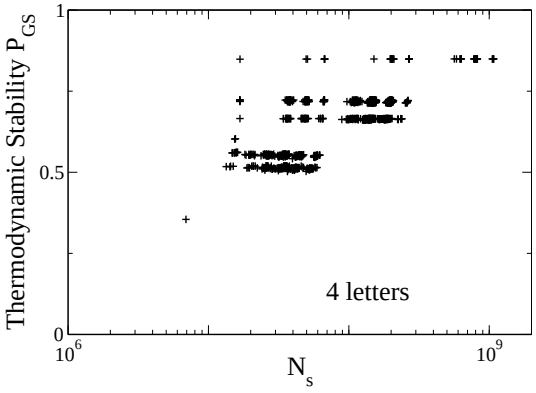

FIG. 6: Stability [20] versus designability N_s (in logarithmic scale) for 24-base RNA sequences constructed with four types of bases. We find no significant correlation between designibality and stability for four letters.

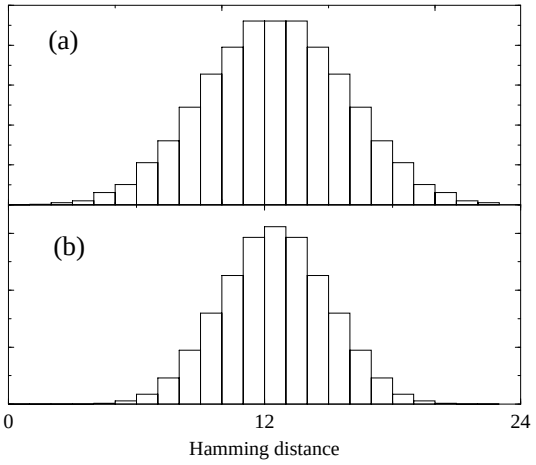

FIG. 7: For some given 24-nucleotide two letter sequence σ_1 we plot (a) a histogram of the distances to all two-letter sequences with the same ground state structure, and (b) a histogram of the distances to all two-letter sequences. Histogram (b) is independent of the choice of σ_1. Histogram (a) is also roughly independent of sequence σ_1 provided its ground-state structure is highly designable.

similar. We trace this difference between two and four or six-letter sequences to the likelihood of competing repaired states. For two letters, the correlation between designability and stability (as well as the nontrivial distribution of designabilities) arises primarily from competing re-paired states. Four and six-letter sequences have far fewer competing re-paired states and hence do not demonstrate significant correlation between designability and stability.

C. Neutral Networks

Finally we consider the "neutral network" of RNA sequences which fold to a particular structure. The connectivity within a network and the shortest distance between networks has drawn considerable attention with respect to the evolvability of RNA structures[8, 22]. In our model, the network of sequences which fold to a particular structure is truly "neutral" in that all sequences have the same ground-state free energy $F_{\rm GS}$, albeit with different stabilities because of repairing. (This contrasts with protein solvation models in which, independent of competing structures, there is typically an energy hierarchy of sequences for each structure, determined by the match between hydrophobicity and surface-exposure pattern[14].) In our model, RNA sequences that fold to a given structure form, in general, a percolating and non-compact network in sequence space. In particular, a histogram of the distances between sequences folding to the same highly designable structure is actually broader than a histogram of the distances between *all* sequences (Fig. 7)[23]. In this respect, the RNA model differs considerably from protein models.

IV. CONCLUSIONS

To conclude, in this paper we developed and studied a minimalist base-stacking model of RNA secondary structure. We found that sequences constructed with four or six types of bases typically have fewer competing excited states, and, consequently, have greater ground-state stability, compared to sequences constructed with two types -of bases. At the same time, the fraction of sequences with highly stacked ground states is much smaller for four-letter sequences than for two, and much smaller for six letters than for four. It is tempting to speculate that four letters optimizes the stability of structures while maintaining a reasonable probability that a random sequence folds into a highly stacked structure. If, as has been postulated, early life was indeed RNA based and double-stranded DNA came later in evolution, our observations might plausibly bear on nature's choice of four letters for the genetic code.

Acknowledgments

We thank David Moroz for useful discussions and suggestions.

[1] The RNA World (Monograph/Cold Spring Harbor Laboratory, No 2), Ed. Raymond F. Gesteland and John F. Atkins, (1993).
[2] See, for example, A.D. Ellington, Current Biology **4**, 427 (1994).
[3] B.M.R. Stadler, P.F. Stadler, G.P. Wagner, and W. Fontana, J. Theo. Bio., **213**, 241-274 (2001). L.F. Landweber, Trends in Ecology and Evolution **14**, 353 (1999).
[4] I. Tinoco and C. Bustamante, J. Mol. Biol. **293**, 271 (1999).
[5] S.Y. Le and M. Zuker, J. Mol. Bio. **216**, 729 (1990).
[6] M. Zuker and D. Sankoff, Bull. Math. Biol. **46**, 591 (1984).
[7] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster, Monatshefte f. Chemie **125**, 167 (1994).
[8] W. Fontana, D.A.M. Konings, P.F. Stadler, and P. Schuster, Biopolymers **33**, 1389 (1993).
[9] J.S. McCaskill, Biopolymers **29**, 1105 (1990).
[10] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999); R. Bundschuh and T. Hwa, Phys. Rev. E **65**,

031903 (2002).
[11] We use a narrow range of stack numbers to emphasize the dependence of stability on alphabet size. For a wider range of stack numbers, the increase in stability with stack numbers can obscure the dependence on alphabet sizes.
[12] H.S. Chan and K.A. Dill, J. Chem. Phys. **92** 3118 (1990). E. Shakhnovich and A. Gheutin, J. Chem. Phys. **93** 5967 (1990).
[13] H. Li, R. Helling, C. Tang, and N.S. Wingreen, Science **273**, 666 (1996).
[14] H. Li, C. Tang, and N.S. Wingreen, Proc. Natl. Acad. Sci. USA **95**, 4987 (1998).
[15] R. Mélin, H. Li, N.S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999); S. Govindarajan and R.A. Goldstein, Biopolymers **36**, 43 (1995).
[16] E.L. Kussell and E.I. Shakhnovich, Phys. Rev. Lett. **83**, 4437 (1999).
[17] H. Li, N.S. Wingreen, and C. Tang, Proteins **49**, 403 (2002).
[18] J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Proteins **47**, 506 (2002).
[19] For each structure, we generate a random sample of sequences that are compatible with the structure and calculate the fraction of such sequences that have this structure as the ground state. We multiply the total number of compatible sequences by this fraction to obtain the designability.
[20] In Fig. 4, we have kept only structures that are highly stacked, in particular, those that have six or fewer unpaired bases. Our results do not depend sensitively on this cutoff.
[21] In Fig. 4, we plot the 90-th percentile of greatest stability, $P_{90\%}$, rather than average stability of sequences folding to a structure. Since sequences folding to a structure have, in general, a wide range of stabilities, the two can be quite different. Average stability shows a similar, but less pronounced correlation with designability than $P_{90\%}$.
[22] W. Fontana and P. Schuster, J. Theor. Biol. **194**, 491 (1998).
[23] A "Hamming" distance between RNA sequences can be defined as a distance of 1 between non-identical bases at a site and 0 between identical bases.